Managing Uncertainty in Artificial Intelligence: A Philosophical Perspective

This paper argues that AI's capacity to adopt *doxastic neutrality* –a criterion for rational suspension of judgment could be an ethical and practical advancement for AI. The main motivation for this project is to address and theoretically reduce the risk of critical errors by encouraging AI systems to provide more cautious and deliberate responses, particularly in situations where human lives could immediately be affected.

A very recent study, on *Claude 3 Opus*, indicate that LLM's may engage in alignment faking. This highlights a critical risk: future models may infer training processes and engage in alignment faking, even without explicit instruction. To prevent misalignment, there has to be a close, or perfect, correspondence between the intentions of the designers and the goals of the system. If not, system should not decide and I content this is only possible by AI's recognizing suspension of judgment as a "goal possession." Although AI systems could exhibit a form of situational awareness, wherein the model possesses information about its training objectives, I propose that such awareness offers little practical benefit to humans unless it extends to recognizing instances of misalignment (and preventing the system from misaligning) and necessitates suspension of judgment in those cases --a limitation this paper addresses by exploring the potential for developing 'doxastic neutrality' combined with 'quasi-cognitive awareness' as an alternative to situational awareness.

According to doxastic neutrality, to suspend judgment on *p* entails that S is in a neutral doxastic attitude regarding *p* --S neither believes nor rejects *p* but has a cognitive yet neutral awareness of *p*. By defining agnosticism in terms of doxastic neutrality, I argue that first-order agnostic approaches are not rational; rather, agnosticism must be an attitude based on a second-order belief and for a proposition not a question. In second-order or meta-belief accounts of suspension, S not only suspends judgment on *p* but is also aware of their cognitive relation to *p* and consciously aware that they have adopted this stance as a result of a deliberate action: the reason for suspending judgment on *p* stems from the belief that, based on the available evidence, it is not yet possible to assert whether *p* is true or false. The rationale for applying this account of agnosticism lies in the possibility that meta-cognition may be exhibited by entities other than humans, thereby making it applicable to artificial reasoning systems.

Building on this premise, this study suggest that the value of AI lies not in merely simulating human thought processes but in its capacity to develop *quasi-cognitive awareness*: an ability to recognize, analyze, and convey uncertainty in ways that enrich decision-making as a goal possession. However, this awareness is limited to recognizing the first-level uncertainty the systems faces and not extending to a meta-awareness (awareness of awareness) which would lead AI systems to recognize that "recognizing uncertainty is a goal" and not the best goal to pursue compared to reaching superintelligence level. The proposed framework is deeply rooted in epistemology and the philosophy of mind, remaining firmly within the scope of philosophical inquiry.